



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Neurogenetics II

Citation for published version:

Wright, AF 2005, 'Neurogenetics II: complex disorders', *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. 5, pp. 623-31. <https://doi.org/10.1136/jnnp.2004.047704>

Digital Object Identifier (DOI):

[10.1136/jnnp.2004.047704](https://doi.org/10.1136/jnnp.2004.047704)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Journal of Neurology, Neurosurgery & Psychiatry

Publisher Rights Statement:

available via europepmc link

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



REVIEW

Neurogenetics II: complex disorders

A F Wright

J Neurol Neurosurg Psychiatry 2005;76:623–631. doi: 10.1136/jnnp.2004.047704

The genetic analysis of common neurological disorders will be a difficult and protracted endeavour. Genetics is only one of many disciplines that will be required but it has already thrown considerable light on the aetiology of several major neurological disorders through the analysis of rare inherited subgroups. The identification of individual susceptibility genes with variants of smaller effect will be more difficult but there is no sharp demarcation between large and small genetic effects, so that many new and important insights will emerge using existing and new technologies. The availability of improved neuroimaging, better animal models of disease and new genetic tools, such as high-throughput gene chips, expression microarrays and proteomics, are extending the range of traditional genetic mapping tools. Finally, an understanding of the genetic and epigenetic mechanisms that restrain the differentiation and integration of human neural stem cells into mature neuronal networks could have a major impact on clinical practice. These approaches will be illustrated in the context of Alzheimer disease, Parkinson disease and synucleinopathies, tauopathies, amyotrophic lateral sclerosis and stroke.

between one third and two thirds of the typical variation in a complex trait, with at least some effect on reproductive fitness, results from rare variants with adverse effects.⁴ The remainder is due to common variants, many of them with opposite effects on different traits (some beneficial, others detrimental) allowing them to be maintained in the population. The motivation for finding common variants is currently greater than for finding rare variants, for three main reasons. First, they provide potential mechanistic insights; second, they are easier to identify than rare variants; third, and most important, they may be of public health importance and allow identification of subpopulations at increased risk of disease.⁵

The success in finding both common and rare genetic variants influencing susceptibility to Alzheimer disease (see below) shows that both CD/CV and CD/RV models are “correct”, but it is a matter of debate as to which will provide the most useful insights. This is perhaps the biggest issue at stake, since the majority of complex traits are polygenic—resulting from the combined action of many different genes, in combination with often proportionately greater environmental effects. In addition, recent evidence suggests that interaction effects—gene-gene and gene-environment—are common, even in experimental organisms where genotype and environment are well controlled.⁶

The methods currently being used to unravel the genetics of common neurological disorders, such as Alzheimer disease and stroke, are essentially the same as those used in the early phase of the Human Genome Project, namely low resolution genetic mapping by linkage analysis in families with multiple affected individuals, followed by high resolution mapping using case-control association studies. However, increasing emphasis is being placed on the latter, fuelled by technological advances using single nucleotide polymorphism (SNP) chips (see Appendix A). However, the large scale use of candidate gene association studies has led to a serious problem, with many unreplicated and, in many cases, spurious associations being published. As an example, out of 127 candidate gene

The sequencing of the human genome signalled a major shift in the Human Genome Project from gene discovery in monogenic disorders towards the “post genome challenge” of gene characterisation and the genetic analysis of complex disorders. This change was largely driven by the increasing facility of gene identification, which led to the identification of >1200 predominantly Mendelian disease genes. An important conceptual development was the common disease/common variant (CD/CV) hypothesis as a model for complex disorders^{1 2} (see Appendix A, Human genetic variation). This model proposed that the genetic basis of common, genetically complex disorders is principally due to genetic variants that are common in the population. In contrast, the common disease/rare variant (CD/RV) model argued that in complex disorders there is a significant contribution from rare variants, which include most of those with the most significant individual effects.^{3 4} Although the debate continues, the heritability of a complex trait almost certainly results from both common and rare variants. One estimate, based on more than two decades of research on such traits in the experimentally tractable organism *Drosophila melanogaster*, suggested that

Correspondence to:
A F Wright, MRC Human
Genetics Unit, Western
General Hospital, Crewe
Road, Edinburgh EH4
2XU, UK; alan.wright@
hgu.mrc.ac.uk

Received 15 June 2004
Revised version received
19 October 2004
Accepted
8 November 2004

Abbreviations: ALS, amyotrophic lateral sclerosis; APP, amyloid β protein precursor; ASPs, affected sib pairs; CD/CV, common disease/common variant; CD/RV, common disease/rare variant; CSF, cerebral spinal fluid; FAD, familial Alzheimer disease; FPD, familial Parkinson disease; FTD, fronto-temporal lobe dementia; IBD, identical by descent; LD, linkage disequilibrium; MS, multiple sclerosis; NFD, neurofibrillary degeneration; PHF, paired helical filament; QT, quantitative trait; SN, substantia nigra; SNP, single nucleotide polymorphism; SOD1, superoxide dismutase 1

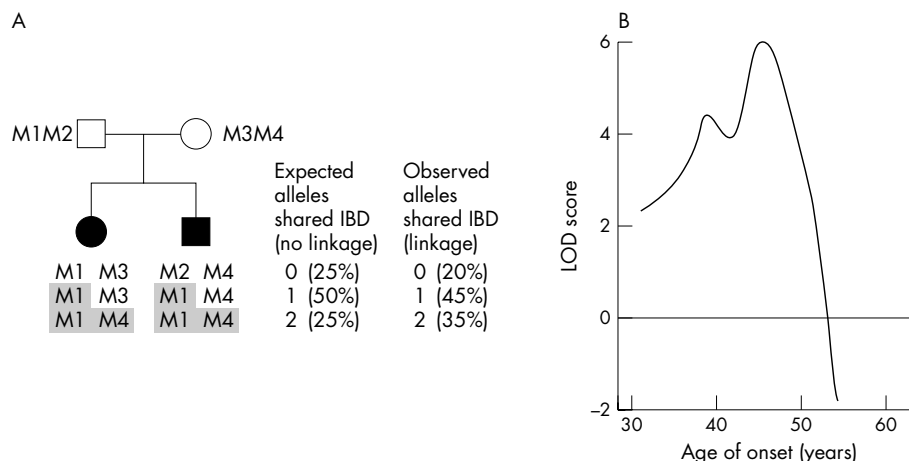


Figure 1 The principles underlying genetic linkage analysis of a common neurological disorder. (A) Affected sib pairs (ASPs) are first genotyped for several hundred genetic markers. For late onset disorders the parents are usually unavailable, but allele sharing between ASPs can still be inferred. In the absence of linkage, the extent of genetic marker allele (M1–4) sharing between ASPs is zero, one, or two alleles shared (identical by descent or IBD) as a result of common ancestry. This is expected to occur 25% (no sharing), 50% (one allele shared), or 25% (both alleles shared) of the time, under the null hypothesis of no linkage (shared alleles are shaded in grey). In the presence of linkage, there is an increase in allele sharing over the null hypothesis, as shown. This method is robust when the precise mode of inheritance is unknown. Large numbers of ASPs (for example, 500–1000) are often required to accumulate significant evidence of linkage to a common disorder. Affected individuals are shown as filled boxes (males) or circles (females) and unaffected individuals are unfilled. (B) Linkage to a late onset disorder may vary according to age of onset. These LOD score data are from Hall *et al.*⁶² who reported stronger evidence for linkage to the *BRCA1* gene on chromosome 17 in early compared with late onset familial breast cancer families. The LOD score is the \log_{10} of the likelihood ratio, in which the probability of linkage at a specified genetic distance is compared with the likelihood under the null hypothesis of no linkage. A likelihood ratio of 1000 (LOD score 3) corresponds to significant linkage (p -value ~ 0.05) for a monogenic disorder while a LOD score just over 3 is significant for a complex disorder. A LOD score just over 3 is significant for a complex disorder. The cumulative LOD scores are shown for all families in which the mean age of onset is less than or equal to the age shown on the x axis.

associations with Alzheimer disease reported in a single year, only three were found to have been replicated in three or more independent studies.⁸

A number of principles which have emerged to guide researchers through the maze of complex genetic disorders are discussed below.

GUIDING PRINCIPLES

Large sample sizes

Most individual genetic effects on complex traits or diseases are small, emphasising the need for large sample sizes to reliably detect them.⁹ Very few genes are capable of exerting large effects, but many genes can exert small marginal effects. A widely accepted model for the distribution of effect sizes of genetic variants influencing complex traits is an L shaped distribution—many genes with variants showing small and peripheral effects on disease (both rare and common) and a smaller number with variants showing moderate to large effects (which tend to be rare).⁴ The effect of individual variants will therefore often be obscured by those of other genes and by large environmental and interaction effects.

Quantitatively varying intermediate disease endpoints

Quantitative traits (QTs) which influence disease risk are used whenever possible to increase study power. In a recent review, it was commented that “studies using a single clinical endpoint are akin to a shot at the moon”, and compare unfavourably with studies focusing on genetically and physiologically simpler intermediate traits.⁷ All individuals with QT information are informative in genetic mapping studies, in contrast to studies focusing on disease, where most of the power comes from the comparatively few affected individuals. It has been difficult to find useful QTs in neurological disorders, compared with cardiovascular or metabolic diseases. The use of disease age of onset or severity, plasma amyloid β_{42} in Alzheimer disease, well

validated questionnaires, and structural brain imaging may facilitate this process.

Ascertainment strategies

It is relatively easy to study “typical” patients with disease, but other ascertainment schemes are more powerful. Families of individuals with complex disorders do not generally have multiple affected members, since the incidence in relatives declines exponentially with decreasing relationship to the proband, as expected under a polygenic model. The identification of individuals at the extremes of the QT distribution is helpful in contributing to study power. Extreme individuals may show large genetic effects, without necessarily developing overt disease (for example, because they lack other risk factors). Screening of large samples may therefore be required to detect such extreme individuals. For example, a study of personality traits targeted 88 000 individuals to fill in a postal questionnaire, which identified over 34 000 sib pairs including many with extreme or contrasting trait values. A genetic linkage analysis of extreme or discordant sib pairs led to the identification of several significant linkage peaks.¹⁰ Similarly, the ascertainment of rare individuals with early onset Parkinson disease was necessary for the identification of a major gene (*DJ-1*) causing this disorder.¹¹

Genetic linkage and case-control association designs

These two methods form the core of the genetic mapping effort. Linkage analysis is carried out using extended or small nuclear families (for example, affected sib pairs) (fig 1). The term “genetic linkage” refers to the finding of an association between disease and genetic marker within each of a series of families containing two or more affected individuals, after carrying out a whole genome scan. The latter involves genotyping many “genetic markers”—variant sites unrelated to gene function which show common variation in the general population—situated at regular intervals throughout the genome (see Appendix B, Genetic linkage and association

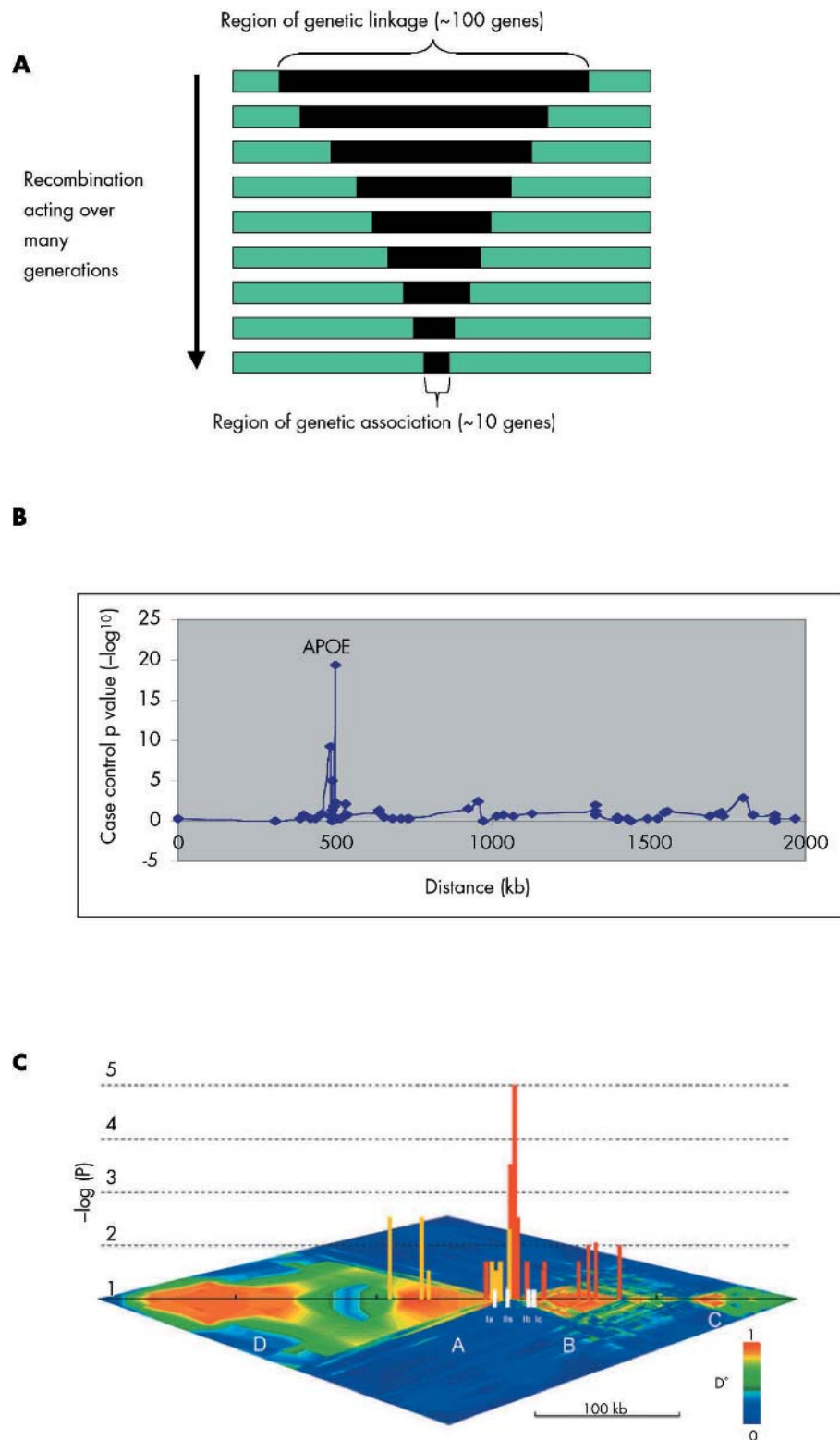


Figure 2 (A) Genetic fine mapping using association studies. Genomic regions shared IBD are shown in black. (B) Allelic association between Alzheimer disease and SNPs across the region flanking the *APOE* gene on chromosome 19.⁶³ As cases are only very distantly related, the region of shared genome is greatest within ~40 kb of *APOE* and falls off steeply beyond this, implying a very dense genome scan would have been required to identify *APOE*. The association was originally identified by genetic linkage to a broad region on chromosome 19q followed by candidate gene association studies. (C) Linkage disequilibrium (LD) map of the *DPP10* gene associated with asthma⁶⁴ showing regions that are associated or in LD with each other as a result of non-random association between pairs of markers. The chromosomal region runs from left to right on the x axis at the bottom of the figure. The strength of association to asthma (red) and the QT immunoglobulin E levels (\log_e IgE) (yellow) is plotted as $-\log(P)$ against position. The markers showing strongest association correspond to the highest peaks. The graph is superimposed on the distribution of LD between markers (measured as D'), which are colour coded and plotted at the marker locations with red (high LD) and dark blue (low LD) at opposite ends of the scale. The four initial exons of the causal *DPP10* gene are shown as white bars.

analyses). If successful, genetic linkage can identify a large genomic region, often containing several hundred genes, in which the disease gene is sought. Linkage disequilibrium (LD) implies non-random association between a pair of markers. This is common for markers that are located close to one another and can occur for several different reasons. The presence of LD between SNP markers makes it possible to infer the location of a disease gene that is in LD with a genotyped SNP. Fine mapping is carried out using the more familiar case-control association study design (fig 2) in which excess marker sharing is sought within cases compared to controls, following a more dense marker genotyping effort within the identified region. In fine mapping, a broad region of genetic linkage, often containing about 100 genes, is narrowed by carrying out dense SNP marker genotyping across the region in cases and controls. This identifies small shared ancestral regions that are associated either with cases or controls. Since the common ancestor is remote, genomic regions that are shared IBD (shown in black in fig 2) become progressively smaller over successive generations as a result of recombination. The number of genes in the identified region of association now contains a finite number of candidate genes which can be analysed for sequence variation.

Choice of study population

Modern urban populations are often extremely diverse and are far from ideal for gene mapping studies because of genetic heterogeneity.¹² However, there is a trade off between obtaining large well characterised study cohorts, which are generally available in urban contexts, and smaller but more homogeneous cohorts from less diverse population groups. The Icelandic population was chosen to study complex diseases to minimise both genetic and environmental heterogeneity, which led to the discovery of several susceptibility genes, including the *PDE4D* gene in stroke (see below).¹³

Choice of research strategy

The research strategy should be specifically designed to answer the question posed. If the aim is to identify common variants with predominantly small genetic effects on a categorical endpoint, such as disease, a broadly based candidate gene screening approach may be appropriate, using common genetic variants (SNPs) (see Appendix A, Human genetic variation) and a hierarchical case-control strategy. For example, a moderate number of cases (for example, $n = 500$) and matched controls could be systematically screened for association between disease and candidate gene SNPs (at an appropriate density per gene). Candidate genes could be selected on the basis of (a) expression within a tissue of interest (for example, hippocampus or substantia nigra), (b) functional criteria, such as membership of a known disease pathway, or (c) localisation to an implicated chromosomal region, on the basis of previous genetic linkage studies. Positive associations could then be followed up using an independent and preferably larger cohort, to eliminate false positives. Alternatively, if the goal is to identify rarer genetic variants of intermediate effect, the strategy could be quite different. A genetic linkage analysis using a large set of families segregating for a QT or disease would be an appropriate initial strategy, as used to identify the chromosomal locations of the *APOE*, α T-catenin, and *GSTO1* genes (see below). Fine mapping could follow using a case-control association study, and a dense set of SNPs confined to the implicated region(s) (fig 2).

Nature of disease susceptibility variants

Susceptibility genes in complex diseases are often expressed in a wide range of tissues and may contain only subtle

variants or combinations of variants, some or all of which lie outside protein coding sites. This makes identification of susceptibility variants difficult. Overall, about 5% of the human genome has functional significance and so is potentially involved in disease.¹⁴ About 1.5% of the genome contains the protein or RNA coding regions of the 20 000–30 000 human genes, in which lie an estimated 20 000 coding or cSNPs.⁵ These represent an important initial target for whole genome association studies. Firstly, they are more likely to influence disease than non-coding SNPs and, secondly, a genome scan could be carried out using substantially fewer markers than the estimated 600 000–1 000 000 non-coding SNPs required to provide coverage of the entire genome.⁵ A further 1% of the human genome lies within genes and is transcribed but is not translated into protein. Finally, an additional 2.5% of the genome lies outside of the genes altogether but is conserved across species, suggesting that these regions also have functional importance. Proving that such subtle non-coding variants influence a complex disease is difficult. In monogenic disorders, the situation is quite different, with 99% of mutations occurring in protein coding or splice sites, and only 1% within non-coding regulatory regions.¹⁵ The best evidence that a gene influences disease susceptibility comes from the identification of several different genetic variants within its coding or splice sites in different affected (or extreme QT value) individuals, coupled with the demonstration that variants affect gene function and show relevant tissue expression.

APPLICATIONS TO CLINICAL NEUROLOGY

Alzheimer disease

Alzheimer disease provides an excellent paradigm for the genetic basis of a complex disorder, with contributions from both common modifier genes and rare variants of large effect.¹⁶ Heritability estimates in Alzheimer disease are in the region of 60%,¹⁷ suggesting that genetic variation plays a significant role in the disease process. However, the major insights into disease mechanisms to date have come from mutations in genes that are so rare that they make essentially no contribution to the heritability of the disease as a whole.

One of the best paradigms for the CD/CV hypothesis was the discovery of common variants in the *APOE* gene which influence susceptibility to Alzheimer disease. There are three common *APOE* alleles (E2, E3, E4) in human populations, resulting from differences at two amino acid residues (residues 112, 158).¹⁸ Associations between the E4 allele, which is present in about one third of Caucasians, and Alzheimer disease have been widely confirmed, but associations have also been found with several other disorders—the Lewy body variant of Alzheimer disease, Parkinson disease, susceptibility to herpes simplex virus infection, poor recovery from head injury, intracerebral haemorrhage, and elective cardiac bypass surgery.¹⁹ A protective effect of the E2 allele in Alzheimer disease has also been reported. *APOE* is the primary cholesterol transporter in the brain and is a component of both amyloid (senile) plaques and neurofibrillary tangles. The mechanism for the effects of *APOE* isoforms on brain damage and dementia is unclear, although transgenic ApoE deficient mice (*ApoE*^{−/−}) engineered to express a human *APOE* E4 allele showed age related spatial learning and memory defects, in contrast to *ApoE*^{−/−} controls or mice carrying the E3 allele.²⁰ Lipid carrying apoE3 binds amyloid β (A β) peptide, the major constituent of amyloid plaques, with 20-fold higher affinity than lipidated apoE4, which may enhance the clearance of A β .²¹ The close relationship between *APOE* and Alzheimer disease risk is highlighted by the finding that transgenic mice overexpressing familial Alzheimer disease mutations on an *ApoE*^{−/−} null

background show very little A β amyloid deposition, compared with those on a normal (wildtype) ApoE^{+/+} background.²² This suggests that *APOE* is essential for A β deposition in transgenic models of familial Alzheimer disease (FAD). It remains unclear whether this effect is mediated by increased formation or decreased clearance of A β amyloid.

The effect of the *APOE* E4 allele is dosage dependent, so that carriers of a single E4 copy have a twofold increased risk of Alzheimer disease compared with a fivefold risk for homozygotes with two copies. The E4 allele appears to be a disease modifier, exerting its effect on disease risk by influencing age of onset in both Alzheimer disease and Parkinson disease, rather than disease risk per se. Despite the relatively large effects of these variants, the use of *APOE* genotype information in disease prediction remains limited, since its diagnostic sensitivity is only 0.65 and specificity 0.68, compared with clinical diagnosis, which has a reported sensitivity of 0.93 and specificity of 0.55.²³

A number of Alzheimer disease modifier loci have recently been proposed, none of which have yet been consistently replicated, but they illustrate some of the approaches taken and difficulties encountered. The glutathione-S-transferase (*GSTO1*) gene was proposed to be a determinant of age of onset, here used as a QT, in both Alzheimer disease and Parkinson disease.²⁴ *GSTO1* is widely expressed and is thought to be concerned with the biotransformation of compounds such as free radicals and interleukin-1 β . The gene was identified by narrowing the number of genes in the large region of chromosome 10 implicated by linkage analysis from several hundred genes to only four, on the basis that only these genes showed altered expression in the hippocampus of Alzheimer disease compared with control subjects. This is an interesting but potentially misleading assumption. Using a case-control strategy, and large sample sizes, the authors found a significant association with one of the three genes, *GSTO1*.²⁴ One of the common variants analysed, SNP7, was associated with the substitution of aspartic acid for alanine at residue 140 (Ala140Asp) in the *GSTO1* product. However, since about 90% of the population carry one or two copies of this early onset "risk" allele (Ala140), it remains unclear how much of the original linkage signal is explained by this (and the associated SNP9) variant, or how useful the resultant mechanistic insights will be.

The identification of another proposed genetic modifier in Alzheimer disease followed the discovery of an association between the insulin degrading enzyme (*IDE*) gene and Alzheimer disease itself,^{25–26} age at onset in both Alzheimer and Parkinson disease,²⁷ and plasma amyloid A β ₄₂, sometimes used as a QT risk factor for Alzheimer disease.^{28–30} The A β ₄₂ peptide is a secreted cleavage product of the amyloid β protein precursor (APP), which is strongly expressed in brain and cerebral spinal fluid (CSF). A β ₄₂ is present in CSF at 50 times its concentration in plasma, but, in a longitudinal study, individuals who developed Alzheimer disease showed higher levels of plasma A β ₄₂, suggesting its use as a surrogate for brain A β ₄₂ production. Plasma A β ₄₂ is elevated in individuals with familial late onset Alzheimer disease, in early onset FAD, and in Down syndrome (since the *APP* gene is carried on chromosome 21). It remains unclear which variants in or close to the *IDE* gene are directly concerned with Alzheimer disease risk, age of onset, and plasma A β ₄₂ levels. *IDE* is an interesting candidate gene since it has been shown to regulate A β ₄₂ levels in brain neurons and microglial cells.^{29–30} Increased degradation of A β ₄₂ by transgenic mice overexpressing *IDE* or another A β -degrading protease, neprilysin, slows A β ₄₂ deposition and reduces Alzheimer-like pathology in mouse models of FAD.³¹

The most significant advances in the genetics of Alzheimer disease and Parkinson disease to date have come not from

the identification of the common variants discussed above, but from the study of genes which have virtually no role in common forms of these disorders. Mutations in three genes account for about half of all cases of FAD,³² which is an extremely rare disease, with fewer than 200 confirmed FAD families worldwide, compared with an estimated 4–5 million Alzheimer disease individuals in the USA alone.³³ FAD is clinically and pathologically indistinguishable from Alzheimer disease except for age of onset. The most common cause is a mutation in the presenilin-1 (*PS1*) gene, which is found in about half of all FAD families. Mutations in the related presenilin-2 (*PS2*) gene and in the *APP* gene account for <1% and <5% of FAD families, respectively.³² Mutations in all three genes give rise to increased A β ₄₂ formation since the presenilins form part of a protein complex concerned with the processing and release of the neurotoxic A β ₄₂ peptide from APP.¹⁶ Mutations in the *APP* and *PS1* genes give rise to a fully penetrant autosomal dominant disorder with onset in the age range 35–55 years, while *PS2* mutations are more variable, often showing later onset (age range 40–85 years) and occasional non-penetrance.

The importance of these rare mutations lies in the identification of a pathogenetic pathway, involving the endoproteolytic cleavage of the transmembrane APP protein by the enzymes BACE1 and the γ -secretase complex.¹⁶ The common factor in Alzheimer disease arising from Down syndrome and mutations in the *APP*, *PS1*, and *PS2* genes is an excess production of the neurotoxic A β ₄₂ peptide or an increased ratio of A β ₄₂ to the less toxic A β ₄₀ peptide. Paradoxically, the pathogenetic sequence in the transition from old age through mild cognitive impairment to Alzheimer disease emphasises the role of neurofibrillary degeneration (NFD), associated with paired helical filament (PHF)-tau deposition, rather than amyloid plaque formation.^{34–35} Amyloid deposits are deposited randomly throughout the entire cerebral cortex, and tend to appear subsequent to NFD and PHF-tau deposits in any one region. NFD progresses hierarchically along specific neuronal pathways (starting in the trans-entorhinal cortex and progressing to the temporal cortex), suggesting a specific vulnerability in these pathways. It has been suggested that this vulnerability may be enhanced in the presence of increased A β ₄₂ formation, which can result from genetic mutations or environmental events such as head injury or stroke. There is an apparent progression in the extent of both NFD and amyloid deposits from normal ageing to Alzheimer disease. For example, in one study, 100% of individuals over age 75 showed NFD in the hippocampus, often in the absence of amyloid plaques or dementia, whereas those with Alzheimer disease (by definition) also have both significant neuronal loss and amyloid plaques.³⁴

TAUOPATHIES

The discovery of mutations in the *Tau* gene in a subset of patients with fronto-temporal lobe dementia (FTD) linked to chromosome 17 (FTDP-17) throws further light on Alzheimer disease mechanisms.³⁶ FTD is an early onset (<65 years) disorder associated with prominent frontal lobe symptoms, such as behavioural disinhibition, with fronto-temporal atrophy due to neuronal loss, spongiform degeneration, and gliosis, sometimes extending to the substantia nigra (SN), amygdala, and spinal cord. Clinical presentation can be accordingly varied. There are no amyloid or Lewy bodies and a small proportion of patients have *Tau* gene mutations.³⁷ Tau is a phosphoprotein expressed in peripheral and central nervous systems, predominantly in neurons, where it is associated with axons and concerned with the microtubule binding and assembly that is necessary for axoplasmic transport.³⁷ Hyperphosphorylated Tau deposits are associated

with PHF and the NFD found in Alzheimer disease. In FTDP-17, both loss of function mutations and mis-expression of the *Tau* gene, which is normally processed into different isoforms, are found. The precise disease sequence and mechanism remains unclear, but amyloid A β_{42} overexpression appears to exacerbate Tau pathology. One possibility is that APP mis-processing in Alzheimer disease leads to post-translational modification of the Tau protein and subsequent neurodegeneration. The observation that amyloid deposition follows rather than precedes Tau mis-processing could however also be explained by the proposal that A β_{42} neurotoxicity results from formation of the more toxic soluble protofibrils rather than the later appearing insoluble fibrillar aggregates.³⁸

Parkinson disease and synucleinopathies

The presence of neuronal loss and insoluble aggregates of α -synuclein, called Lewy bodies, in the SN are the major pathological features of Parkinson disease.³⁹ Surprisingly, the prevalence of SN Lewy bodies in the general population is ten times greater than the prevalence of Parkinson disease, but there appears to be a threshold, so that those with SN neuronal loss exceeding about 60% show symptoms of Parkinson disease. This may be because in disorders of protein aggregation, the characteristic aggregates are actually protective but when present in large numbers are indicative of a more sinister underlying process or extent of disease. Post mortem studies show that SN cell loss in the normal population follows an exponential distribution, with 4.4% of cells lost per decade.⁴⁰ In contrast, cell loss in Parkinson disease appears to occur ten times faster, at a rate of 45% per decade, with onset about 4–5 years before symptomatic disease.⁴⁰ Lewy bodies are also a prominent feature in other neurological disorders—dementia with Lewy bodies, multiple system atrophy, Down syndrome, and neurodegeneration with brain iron accumulation I.⁴¹ Ten genes have been mapped by genetic linkage to rare monogenic forms of familial Parkinson disease (FPD), four of which have been isolated: the α -synuclein (*SNCA*), ubiquitin C-terminal hydrolase like 1 (*UCH-L1*), parkin (*PRKN*), and *DJ-1* genes.⁴² These have again provided mechanistic insights into common forms of Parkinson disease. Firstly, mutations in the α -synuclein gene result in early onset autosomal dominant FPD.⁴³ Autosomal dominant FPD families showing triplication or duplication of the *SNCA* gene present FPD symptoms in the fourth and fifth decades respectively, implying that overexpression even of normal α -synuclein is sufficient to cause disease. Genetic variability in the *SNCA* promoter region was associated with increased risk of sporadic Parkinson disease. This is consistent with the possibility that, like overexpression of A β_{42} in Alzheimer disease and Down syndrome, increased formation of normal α -synuclein can be disease causing.

Mutation in the *PRKN* gene causes juvenile or early adult (<45 years) onset autosomal recessive PD.⁴⁴ Complete loss of parkin due to homozygous deletion of the *PRKN* gene is associated with severe loss of dopaminergic neurons in the SN and locus coeruleus but a notable absence of Lewy bodies. Some amino acid changing (missense) mutations in *PRKN* do show both Lewy bodies and abnormal tau deposits (NFD), suggesting a possible gain of function. One explanation is that since parkin is an E3 ubiquitin ligase, it is a component of the ubiquitin proteasome system, which may be required to produce Lewy bodies. The ubiquitin protease system is involved with the degradation of misfolded proteins, some of which—such as α -synuclein and perhaps some types of mutant parkin itself—can give rise to aggregation and neurodegeneration. The importance of this pathway is reinforced by the finding of mutations in the *UCHL1* gene,

coding for ubiquitin carboxy-terminal hydrolase L1, one of the most abundant proteins in the brain, in rare autosomal dominant FPD families.⁴⁵ The *UCHL1* enzyme is found in Lewy bodies and is also concerned with protein degradation. *UCHL1* mutations lead to accumulation of α -synuclein in cells and may influence susceptibility to Parkinson disease by altering the balance of ubiquitin hydrolase and ligase activities, both of which are present in *UCHL1*, impairing the degradation of α -synuclein.⁴⁶

DJ-1 is another component of the ubiquitin/proteasome protein degradation pathway which is mutated in a rare autosomal recessive form of early onset Parkinson disease.¹¹ The gene was identified by genetic linkage analysis in a large inbred Dutch community in which the mutant gene appeared to be more common as a result of a founder effect and cultural isolation of this population. Since both Parkin and DJ-1 are components of the ubiquitin proteasome pathway, and are concerned with the degradation of fibrillogenic proteins within the SN, these rare genes have again identified an important pathogenetic pathway in all forms of Parkinson disease, despite making essentially no contribution to heritability in the common form.

Amyotrophic lateral sclerosis (ALS)

ALS is a progressive disease associated with degeneration of motor neurons in the brain stem and spinal cord. Surviving neurons contain inclusions of neurofilament components and ubiquitin. It is generally sporadic but rare familial forms of ALS occur in about 10% of patients, about 20% of which are associated with missense mutations in the cytoplasmic enzyme Cu/Zn superoxide dismutase 1 (*SOD1*), which is also present in the inclusions.^{47–48} It is unclear whether the disease results from a gain of function, such as protofibril toxicity, or loss of function and oxidative stress. *SOD1* catalyses the dismutation of the superoxide radical to form hydrogen peroxide and oxygen. One possibility is that an oxidising environment (due to reduced *SOD1* activity) causes protein instability, aggregation, and neurotoxicity, since mutant *SOD1* aggregates have been seen under such conditions.

Cerebrovascular disease and stroke

Stroke is a heterogeneous group of ischaemic and, less commonly, haemorrhagic disorders, which are associated with atherosclerosis of large blood vessels or occlusion of small penetrating arteries in the brain. All forms of stroke share common risk factors, including hypertension, hyperlipidaemia, diabetes, and smoking. Family history is an independent risk factor, suggesting that genetic factors may contribute to susceptibility.⁴⁹ Genetic linkage analysis of Icelandic families segregating for stroke provided the initial evidence for a susceptibility gene on chromosome 5. Fine mapping was carried out in a case-control study of 864 affected individuals from the Icelandic population and 908 controls, using 98 markers spanning the implicated chromosomal region. A broad definition of stroke was employed, including both cardiogenic and carotid stroke, and common variants within the phosphodiesterase 4D (*PDE4D*) gene were found to be associated.¹³ The highest risk haplotype (present in 9% of controls) conferred a twofold relative risk. A protective haplotype (present in 21% of controls) was also identified, with a relative risk of 0.7. However, none of the associated variants were present in protein coding or gene splicing regions, suggesting that the identified and/or associated variants affect gene regulation (such as expression level) rather than having a direct functional effect on the protein. Some protein isoforms associated with the risk haplotype may be expressed at a lower level in patients than in controls. The *PDE4D* risk haplotype has an effect that is largely independent of known risk factors. The *PDE4D* gene encodes a cyclic nucleotide phosphodiesterase which

degrades cyclic AMP and regulates signal transduction in a wide variety of cells. One possibility is that *PDE4D* variants cause low cyclic AMP levels, increasing the tendency for proliferation and migration of vascular smooth muscle cells, although similar effects in the immune system are also possible. These findings and their pathogenic significance remain to be confirmed and elucidated.

A similar approach led to the identification of another gene, *ALOX5AP*, coding for 5-lipoxygenase activating protein, in which certain common haplotypes double the risk of both stroke and myocardial infarction.⁵⁰ The initial finding was a suggestive linkage to a region of chromosome 13 in a series of 296 Icelandic families with multiple affected members. A case-control association study was carried out using a high density of markers across the implicated region (containing 40 known genes) which led to the identification of the *ALOX5AP* susceptibility gene. This was confirmed in a UK population, although the associated haplotype was different. The individual or combination of variants associated with disease risk remain to be identified. *ALOX5AP* and 5-lipoxygenase together convert unesterified arachidonic acid to the leukotriene LTA₄, which is further converted to LTB₄ or LTC₄.⁵⁰ These are important proinflammatory mediators which are active in macrophages and leukocytes invading atherosclerotic lesions.

NEW TECHNOLOGIES

Increasing access to powerful new technologies will facilitate the discovery of genetic influences in neurological disorders. Perhaps the most important ones are those concerned with refining the clinical phenotype, such as brain imaging techniques, and developing quantitative intermediate disease endpoints. The goal of reliably defining simpler phenotypes than disease itself, such as carotid intima media thickness, instead of more complex and categorical traits such as stroke, is particularly important. Other enabling technologies are allowing high throughput analysis of genes and their products in health and disease, which is beginning to influence neurological research. The new technologies are discussed below.

Microarrays

High density arrays of DNA sequences, such as SNP alleles or expressed gene sequences (cDNA), can be immobilised on miniaturised grids (chips), in order to perform large scale screening experiments.⁵¹ For example, the messenger RNA (mRNA) from both normal and diseased neurological tissues can be extracted, converted to DNA (cDNA) and labelled prior to hybridisation to the chip, in order to identify genes that are differentially expressed in disease. Alternatively, genomic DNA from an individual could be labelled and hybridised to an SNP chip containing tens or hundreds of thousands of SNP variants, to search for a disease association. Finally, if a candidate gene for a disease has been mapped to a specific genomic region containing a few hundred genes, it may be useful to know which genes from that region are expressed in the diseased region using microarrays.

This technology has been used to investigate neurological disorders.⁵² In one study, cDNA microarrays containing 18 000 genes were hybridised to cDNA from hippocampal CA1 neurons with or without neurofibrillary tangles in Alzheimer and control brains.⁵³ Similarly, prefrontal cortex from schizophrenic versus control brains was screened using arrays containing 7000 genes to detect differences in gene expression, which showed decreased expression of genes regulating presynaptic function.⁵⁴ It is important to confirm changes in gene expression shown by microarray using other methods, such as immunohistochemistry, in situ

hybridisation, or reverse transcription polymerase chain reaction. A final example is the use of microarrays in the transcriptional analysis of brain plaques from multiple sclerosis (MS) samples compared with control brain samples.⁵⁵ This type of study identified osteopontin (*OPN*) gene expression exclusively in MS plaques, which led to the proposal that this proinflammatory molecule is expressed by infiltrating T lymphocytes, microglia, and macrophages, and promotes damage to the myelin sheath as a result of an autoimmune process. Polymorphisms in *OPN* also appear to influence the disease course.^{56 57}

Proteomics

Gene expression profiles provide little information on genetic variation and may give misleading information on the function or expression of their protein products. The proteome, which is the sum of all expressed proteins in a tissue or cell, is regulated at different levels, including synthesis, degradation, and a wide variety of post-translational modifications, such as phosphorylation. The abundance of the mRNA coding for a specific protein may be poorly correlated with protein abundance. However, the variety and different physico-chemical properties of proteins complicates the “protein chip” approach, although the entire yeast proteome has now been arrayed on a chip. Instead, the techniques of two-dimensional gel electrophoresis, in-gel digestion, and peptide identification by microsequencing or mass spectrometry, are together enabling the high throughput analysis and identification of unknown proteins dissected from healthy or diseased tissues. Two-dimensional gel electrophoresis allows the separation of several hundred proteins by molecular size and net charge while techniques such as MALDI-TOF or tandem (q-TOF) mass spectrometry facilitate their identification.⁵⁸ For example, over 300 proteins were identified from subcellular fractions of human frontal cortex using such an approach.⁵⁸ Current limitations include the difficulty of analysing hydrophobic proteins, such as membrane receptors, and the identification of post-translational modifications in a high throughput manner. These techniques however have the potential for refining the analysis of cells and tissues in neurological disorders. Firstly, they can provide critical information on the structure and function of specific proteins, such as disease related post-translational modifications. Secondly, they can provide an overview of the collective changes occurring within a brain region which can help to subdivide and refine molecular subtypes of disease.

Neural stem cells

There is considerable interest in the possibility of inducing resident human neural stem cells, that are known to be present in the subependymal zone and hippocampus, to differentiate into and replace neurons damaged by ischaemia, trauma, or neurodegeneration.^{59–61} This property is retained in the brains of some simpler non-mammalian vertebrates but appears to have been progressively lost with the evolution of increasing brain complexity from amphibians through to rodents and primates. The precise number of human neural stem cells is unknown but <1% of human subependymal cells display the Ki-67 marker that is associated with a capacity for cell division.⁶⁰ In human bone marrow, only about 1 in 10⁶ cells show the properties of haematopoietic stem cells. Human neural stem cells display glial astrocyte but not neuronal markers, although they are able to generate both neuronal and glial cells in culture. It therefore appears that there is an inherent resistance of such cells to undergo neurogenesis in vivo, perhaps because of the need to retain the complex neuronal networks built up by experience and learning. The goal of replacing cells from the temporal or parietal association cortex which are lost in Alzheimer

disease therefore currently seems remote. The more limited goal of understanding the restraints on neural differentiation that limit the neurogenic potential of subependymal neural stem cells in vivo compared with in vitro may well be achievable. This knowledge could ultimately lead to replacement of specific motor or sensory neurons serving less advanced brain functions.

CONCLUSIONS

Genetics is only one of many disciplines that will be required to elucidate disorders like epilepsy and dementia. However, it is a very powerful tool for dissecting such complex phenotypes. Historically, the power of the genetic approach has come from the analysis of relatively simple and rare Mendelian disorders which resemble complex traits or diseases and elucidate key disease mechanisms and pathways. This is well illustrated by the analysis of genes responsible for early onset forms of Alzheimer disease and Parkinson disease. The identification of individual susceptibility genes with variants of smaller effect is proving more difficult. The increased availability of animal models of inherited neurological diseases, and of high throughput gene based technologies, such as microarrays and proteomic analyses, extend the range of traditional genetic tools, such as gene mapping. Finally, an understanding of the genetic and epigenetic mechanisms that restrain the differentiation and integration of human neural stem cells into mature neuronal networks could have a major impact on clinical practice.

The Medical Research Council provided financial support.

Competing interests: none declared

REFERENCES

- Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994;**265**:2037–48.
- Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001;**17**:502–10.
- Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001;**69**:124–37.
- Wright A, Charlesworth B, Rudan I, et al. A polygenic basis for late-onset disease. *Trends Genet* 2003;**19**:97–106.
- Carlson CS, Eberle MA, Kruglyak L, et al. Mapping complex disease loci in whole-genome association studies. *Nature* 2004;**429**:446–52.
- Dilda CL, Mackay TF. The genetic architecture of *Drosophila* sensory bristle number. *Genetics* 2002;**162**:1655–74.
- Kruglyak L, Nickerson DA. Variation is the spice of life. *Nat Genet* 2001;**27**:234–6.
- Bertram L, Tanzi RE. Alzheimer's disease: one disorder, too many genes? *Hum Mol Genet* 2004;**13**(spec no 1):R135–41.
- Dahlman I, Eaves IA, Kosoy R, et al. Parameters for reliable results in genetic association studies in common disease. *Nat Genet* 2002;**30**:149–50.
- Fullerton J, Cubin M, Tiwari H, et al. Linkage analysis of extremely discordant and concordant sibling pairs identifies quantitative-trait loci that influence variation in the human personality trait neuroticism. *Am J Hum Genet* 2003;**72**:879–90.
- Bonifati V, Rizzo P, van Baren MJ, et al. Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science* 2003;**299**:256–9.
- Wright AF, Carothers AD, Pirastu M. Population choice in mapping genes for complex diseases. *Nat Genet* 1999;**23**:397–404.
- Gretarsdottir S, Thorleifsson G, Reynisdottir ST, et al. The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat Genet* 2003;**35**:131–8.
- Bentley DR. Genomes for medicine. *Nature* 2004;**429**:440–5.
- Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003;**33**(suppl):228–37.
- Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* 2002;**297**:353–6.
- Bergem AL, Engedal K, Kringle E. The role of heredity in late-onset Alzheimer disease and vascular dementia. A twin study. *Arch Gen Psychiatry* 1997;**54**:264–70.
- Mahley RW, Rall SC Jr. Apolipoprotein E: far more than a lipid transport protein. *Annu Rev Genomics Hum Genet* 2000;**1**:507–37.
- Mayeux R. Epidemiology of neurodegeneration. *Annu Rev Neurosci* 2003;**26**:81–104.
- Raber J, Wong D, Yu GQ, et al. Apolipoprotein E and cognitive performance. *Nature* 2000;**404**:352–4.
- LaDu MJ, Pederson TM, Frail DE, et al. Purification of apolipoprotein E attenuates isoform-specific binding to beta-amyloid. *J Biol Chem* 1995;**270**:9039–42.
- Bales KR, Verina T, Dodel RC, et al. Lack of apolipoprotein E dramatically reduces amyloid beta-peptide deposition. *Nat Genet* 1997;**17**:263–4.
- Mayeux R, Saunders AM, Shea S, et al. Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. Alzheimer's Disease Centers Consortium on Apolipoprotein E and Alzheimer's Disease. *N Engl J Med* 1998;**338**:506–11.
- Li YJ, Oliveira SA, Xu P, et al. Glutathione S-transferase omega-1 modifies age-at-onset of Alzheimer disease and Parkinson disease. *Hum Mol Genet* 2003;**12**:3259–67.
- Prince JA, Feuk L, Gu HF, et al. Genetic variation in a haplotype block spanning IDE influences Alzheimer disease. *Hum Mutat* 2003;**22**:363–71.
- Ertekin-Taner N, Allen M, Fadale D, et al. Genetic variants in a haplotype block spanning IDE are significantly associated with plasma Abeta42 levels and risk for Alzheimer disease. *Hum Mutat* 2004;**23**:334–42.
- Blomqvist ME, Silburn PA, Buchanan DD, et al. Sequence variation in the proximity of IDE may impact age at onset of both Parkinson disease and Alzheimer disease. *Neurogenetics* 2004;**5**:115–9.
- Ertekin-Taner N, Ronald J, Asahara H, et al. Fine mapping of the alpha-T catenin gene to a quantitative trait locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. *Hum Mol Genet* 2003;**12**:3133–43.
- Farris W, Mansourian S, Chang Y, et al. Insulin-degrading enzyme regulates the levels of insulin, amyloid beta-protein, and the beta-amyloid precursor protein intracellular domain in vivo. *Proc Natl Acad Sci U S A* 2003;**100**:4162–7.
- Farris W, Mansourian S, Leissring MA, et al. Partial loss-of-function mutations in insulin-degrading enzyme that induce diabetes also impair degradation of amyloid beta-protein. *Am J Pathol* 2004;**164**:1425–34.
- Leissring MA, Farris W, Chang AY, et al. Enhanced proteolysis of beta-amyloid in APP transgenic mice prevents plaque formation, secondary pathology, and premature death. *Neuron* 2003;**40**:1087–93.
- St George-Hyslop PH. Molecular genetics of Alzheimer's disease. *Biol Psychiatry* 2000;**47**:183–99.
- Hebert LE, Scherr PA, Bienias JL, et al. Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Arch Neurol* 2003;**60**:1119–22.
- Delacourte A, David JP, Sergeant N, et al. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. *Neurology* 1999;**52**:1158–65.
- Nagy Z, Esiri MM, Jobst KA, et al. Relative roles of plaques and tangles in the dementia of Alzheimer's disease: correlations using three sets of neuropathological criteria. *Dementia* 1995;**6**:21–31.
- Hutton M, Lendon CL, Rizzo P, et al. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* 1998;**393**:702–5.
- Rosso SM, van Swieten JC. New developments in frontotemporal dementia and parkinsonism linked to chromosome 17. *Curr Opin Neurol* 2002;**15**:423–8.
- Caughey B, Lansbury PT. Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. *Annu Rev Neurosci* 2003;**26**:267–98.
- Goldberg MS, Lansbury PT Jr. Is there a cause-and-effect relationship between alpha-synuclein fibrillization and Parkinson's disease? *Nat Cell Biol* 2000;**2**:E115–9.
- Fearnley JM, Lees AJ. Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain* 1991;**114**:2283–301.
- Galvin JE, Lee VM, Trojanowski JQ. Synucleinopathies: clinical and pathological implications. *Arch Neurol* 2001;**58**:186–90.
- Maries E, Doss B, Collier TJ, et al. The role of alpha-synuclein in Parkinson's disease: insights from animal models. *Nat Rev Neurosci* 2003;**4**:727–38.
- Polymeropoulos MH, Lavedan C, Leroy E, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* 1997;**276**:2045–7.
- Kitada T, Asakawa S, Hattori N, et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* 1998;**392**:605–8.
- Leroy E, Boyer R, Auburger G, et al. The ubiquitin pathway in Parkinson's disease. *Nature* 1998;**395**:451–2.
- Liu Y, Fallon L, Lashuel HA, et al. The UCH-L1 gene encodes two opposing enzymatic activities that affect alpha-synuclein degradation and Parkinson's disease susceptibility. *Cell* 2002;**111**:209–18.
- Rosen DR, Siddique T, Patterson D, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 1993;**362**:59–62.
- Brujin LJ, Becher MW, Lee MK, et al. ALS-linked SOD1 mutant G85R mediates damage to astrocytes and promotes rapidly progressive disease with SOD1-containing inclusions. *Neuron* 1997;**18**:327–38.
- Humphries SE, Morgan L. Genetic risk factors for stroke and carotid atherosclerosis: insights into pathophysiology from candidate gene approaches. *Lancet Neurol* 2004;**3**:227–35.
- Helgadottir A, Manolescu A, Thorleifsson G, et al. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat Genet* 2004;**36**:233–9.
- Butte A. The use and analysis of microarray data. *Nat Rev Drug Discov* 2002;**1**:951–60.
- Lockhart DJ, Barlow C. Expressing what's on your mind: DNA arrays and the brain. *Nat Rev Neurosci* 2001;**2**:63–8.
- Ginsberg SD, Hemby SE, Lee VM, et al. Expression profile of transcripts in Alzheimer's disease tangle-bearing CA1 neurons. *Ann Neurol* 2000;**48**:77–87.

- 54 **Mirmics K**, Middleton FA, Marquez A, *et al.* Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron* 2000;**28**:53–67.
- 55 **Steinman L**, Zamvil S. Transcriptional analysis of targets in multiple sclerosis. *Nat Rev Immunol* 2003;**3**:483–92.
- 56 **Caillier S**, Barcellos LF, Baranzini SE, *et al.* Multiple Sclerosis Genetics Group. Osteopontin polymorphisms and disease course in multiple sclerosis. *Genes Immun* 2003;**4**:312–5.
- 57 **Niino M**, Kikuchi S, Fukazawa T, *et al.* Genetic polymorphisms of osteopontin in association with multiple sclerosis in Japanese patients. *J Neuroimmunol* 2003;**136**:125–9.
- 58 **Lubec G**, Krapfenbauer K, Fountoulakis M. Proteomics in brain research: potentials and limitations. *Prog Neurobiol* 2003;**69**:193–211.
- 59 **Eriksson PS**, Perfilieva E, Bjork-Eriksson T, *et al.* Neurogenesis in the adult human hippocampus. *Nat Med* 1998;**4**:1313–7.
- 60 **Sanai N**, Tramontin AD, Quinones-Hinojosa A, *et al.* Unique astrocyte ribbon in adult human brain contains neural stem cells but lacks chain migration. *Nature* 2004;**427**:740–4.
- 61 **Rakic P**. Neuroscience: immigration denied. *Nature* 2004;**427**:685–6.
- 62 **Hall JM**, Lee MK, Newman B, *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990;**250**:1684–9.
- 63 **Martin ER**, Gilbert JR, Lai EH, *et al.* Analysis of association at single nucleotide polymorphisms in the APOE region. *Genomics* 2000;**63**:7–12.
- 64 **Allen M**, Heinzmann A, Noguchi E, *et al.* Positional cloning of a novel gene influencing asthma from chromosome 2q14. *Nat Genet* 2003;**35**:258–63.

APPENDIX A

HUMAN GENETIC VARIATION

Humans are on average 99.9% identical, with one variant base every 1300 base pairs.⁷ Most of the genetic differences between any two individuals consist of SNPs, which are single base changes present in at least 2% of the population (allele frequency >0.01). There are probably over 10 million SNPs and an almost unlimited number of rare variants in the human population.⁷ Most common variants are extremely ancient, pre-dating the divergence of human racial groups >100 000 years ago. They survive in the human genome because the majority are “neutral” in their effects on reproductive fitness. They therefore confer no reproductive advantage or disadvantage. Some common variants have arisen or become common within more recent times (for example, <10 000 years) as a result of selection for some favourable characteristic. In contrast, genetic variants with intermediate or large effects on disease are predicted to be at low population frequency, since they tend to have adverse effects both on disease related traits and on reproductive fitness (which are usually correlated).⁴ Collectively, however, there are many more rare variants than common ones in the human population and these are the ones with large functional effects that contribute most to human Mendelian diseases. It remains to be seen to what extent these rather than common variants provide most insights into common disorders.

APPENDIX B

GENETIC LINKAGE AND ASSOCIATION ANALYSES

A genetic linkage analysis (fig 1A) aims to identify a gene of moderate effect by scanning the genome with several

hundred evenly spaced genetic markers to find one or more that segregates with the trait or disease. An association is first sought between each marker and the trait or disease within each family. The probability of the observed data, assuming either linkage or the null hypothesis of no linkage, are summarised in a LOD score table or graph (fig 1B). In some late onset disorders, the LOD score declines with age of onset, indicating that other factors, such as polygenic or environmental influences, obscure the effect of single genes (fig 1B). Significant evidence of linkage can occur either by chance or because genetic marker and susceptibility gene are adjacent to one another on the same chromosome (true genetic linkage). Different families may have different mutations, but in linkage analysis it is assumed that these occur predominantly within a single gene, and account for much of the variation in disease susceptibility.

A case-control association study compares the frequency of a single SNP marker or more usually a combination of SNPs on a single chromosome (SNP haplotype) in cases and controls (fig 2). An excess of marker alleles or haplotypes in cases compared with controls may occur by chance or as a result of genetic association. A true association occurs when apparently unrelated individuals share a region of the genome as a result of distant common ancestry (fig 2A). In order to identify such regions, a high density of genetic markers is required, which is often restricted to the vicinity of a linkage peak (fig 2B). Association can occur between a disease or QT and genetic marker even if the genetic variant(s) conferring disease susceptibility is not tested directly, provided it is associated with adjacent (tested) markers, due to common ancestry (linkage disequilibrium) (fig 2C). Regions of association between SNP markers are being defined in the HapMap project, which aims to determine the most efficient combinations and density of marker SNPs for disease gene mapping. The aim is to use sufficient well chosen SNPs so that any untested but disease associated SNP will still be detectable in an association study, as a result of its association with adjacent (tested) SNPs (fig 2C).⁵

Genetic association methods work well for fine mapping within a (linkage) defined region, but their use in screening the entire genome for disease susceptibility genes requires very high marker densities—in the region of hundreds of thousands of SNPs, since only a small segment of genome is shared between distantly related individuals (fig 2A). This generates many false positive associations. A second problem is the underlying assumption in association studies that a significant fraction of the variation in disease susceptibility results from not only a single gene, but a single variant within a single gene, making it more restrictive than the linkage approach. It is however a powerful approach for identifying common, small effect variants in large population samples, for example using candidate genes.